Scientometrics (2012) 93:635–644
DOI 10.1007/s11192-012-0766-x

Discovery of factors influencing citation impact based on a soft fuzzy rough set model

Mingyang Wang · Guang Yu · Shuang An · Daren Yu

Received: 2 September 2011 / Published online: 12 May 2012

Abstract In this paper, the machine learning tools were used to identify key features influencing citation impact. Both the papers' external and quality information were considered in constructing papers' feature space. Based on the feature space, the soft fuzzy rough set was used to generate a series of associated feature subsets. Then, the KNN classifier was used to find the feature subset with the best classification performance. The results show that citation impact could be predicted by objectively assessed factors. Both the papers' quality and external features, mainly represented as the reputation of the first author, are contributed to future citation impact.

Keywords Citation impact · Highly cited papers · Soft fuzzy rough set

Introduction

It is well-known that citation distributions are extremely skewed. The vast majority of the scientific papers are never or seldom cited in the subsequent scientific literatures. On the other hand some papers receive an extremely large number of citations. During the last decades, there has been an emerging interest in exploring the reason for this phenomenon.

(i) The citation motivations of quoters were widely discussed (Bornmann and Daniel 2008; Case and Higgins 2000; Hewings et al. 2010; Kim 2004; Laband and Piette

M. Wang (✉)
College of Information and Computer Engineering, Northeast Forestry University,
Harbin 150040, People's Republic of China
e-mail: wangmingyang@nefu.edu.cn

M. Wang · G. Yu
School of Management, Harbin Institute of Technology, Harbin 150001, People's Republic of China

S. An · D. Yu
School of Power Engineering, Harbin Institute of Technology,
Harbin 150001, People's Republic of China

1994; Rong and Martin 2008). But the quoters' citation motivations are subjective to a large extent, which makes it difficult to properly monitor the citation trend of papers.

(ii) Some mathematical-statistical models were established to predict papers' future citation behaviors. Glänzel and Schubert (1995) presented a non-homogeneous birth-process model, and further discussed the statistical reliability of the model (Glänzel 1997). Recently, Burrell presented a series of stochastic models in the presence of obsolescence to predict the future citation pattern of individual papers (Burrell 2001, 2002a, b, 2003). The future citation probabilities for papers: (a) never be cited up to time *t*, and (b) having an arbitrary number of citations up to time *t*, were discussed.

(iii) The bibliometric factors that have influences on papers' citation activities were widely investigated. Features associated with the authors (age, gender, social status, etc.), the papers (collaboration, document type, subject matter, etc.), and the journals (impact factor, etc.) were discussed (Danell 2011; Fu and Aliferis 2010; Levitt and Thelwall 2008; Penas and Willett 2006; Sagi and Yechiam 2008; Xia et al. 2011). However, these features are mainly the external bibliometric features of papers. The factors on papers' *quality*, which could be the kernel features dominate papers' citation activity, are left alone because of lacking an appropriate way to quantify it.

Van Dalen and Henkens (2005) stated that the *quality* of paper could be approximated by the *impact* and *speed* with which knowledge is disseminated in the scientific community. The *impact* of one paper boils down to the number of citations registered by the Web of Science (of ISI). The *speed* with which one paper is disseminated in the scientific community is measured by the timing of the first citation. Recently, we discussed the role of papers' quality on citation impact prediction based on papers published in the field of astronomy and astrophysics in 1980 (Wang et al. 2011). The impact of one paper is expressed by its knowledge diffusion properties in our work. And the speed of one paper is measured by its first-cited properties, including its first-cited age and the citations obtained in its first cited year. By using the technique of multi-classifier system, we found that the papers' quality are contributed to predict papers' future citation impact.

It is well-known that there exist some noises in papers' citation distributions. For instance, some works suggested that the highly-cited papers tend to be published in high impact journals (Van Dalen and Henkens 2001, 2005), but there are also papers present in poorly impact journals. The noisy samples would disturb the citation trend analysis. So the noisy influences must be eliminated. Recently, we found that the soft fuzzy rough set works well to deal with noisy samples (Hu et al. 2010). The soft fuzzy rough set relies on the soft distance, which is distinguished with the statistical minimum distance in other rough set model. It makes it to be well robust to outliers.

In this paper, we identified the key features influencing on papers' citation impact basing on soft fuzzy rough set. The paper set for our experiments were extracted from four different fields. The total citation counts to one paper, which are the past citations from its publication to the year 2010, will serve as a proxy for its future citation impact. The feature space, which was established for each paper, will serve as the set for predictor variables. Not only the external bibliometric features, but also the papers' quality were considered. A feature selection algorithm based on soft fuzzy rough set was used to generate a series of feature subsets. Then, the K-nearest neighbor (KNN) classifier was used to cross-validate the classification accuracy of the data with the feature subsets. And the feature subset with the highest classification accuracy is extracted out, which are the key features influencing on papers' future citation impact.

Data

Four representative journals were chosen from four different fields. All the papers published in these four journals in year 1985 were collected basing on the web version of the Science Citation Index (SCI) produced by ISI. Our goal is to develop a three labels' prediction model rather than a continuous one because error metrics for continuous loss functions are difficult to interpret. The model predicts whether a paper would grow up into highly-cited papers (HCPs), medium-cited papers (MCPs) or low-cited papers (LCPs) within 15 years of publication. A paper set with different kinds of citation impact should be established firstly.

There are various definitions of what counts as a HCPs. Basically two different approaches can be identified, involving absolute or relative thresholds (Aksnes 2003). For the absolute one, a fixed threshold is used as a definition. For example, articles cited more than 500 times are defined as HCPs. The limitation of using a fixed threshold is that the highly cited fields generate a predominance of HCPs. A relative standard is, therefore, often adopted instead. Such a selection method identifies the most highly-cited papers within each field.

In this paper, we used a relative standard. A paper was considered as highly-cited if it has received more than a certain multiple of the citations of the average paper within the scientific subfield. By this we have adjusted for the large differences in citation rates between different subfields. The concept of HCPs is thus based on variable and field specific standards.

Selection criteria

The basic principle in determining the HCPs is that the number of citations received should be more than a certain multiple of the mean citation rate of the particular subfield. Such a method of selection has similarities to the method applied by Glänzel et al. (1995) in a study of HCPs in physics, and the method applied by Aksnes (2003) in a study of HCPs for Norway authors. In addition, the sample identified should be manageable from a practical point of view, meaning that the number of papers should not be too large (or too small) for carrying out the different surveys. Using the references standard described above, we selected a score value of 10. That is, a publication has been considered as highly cited if the number of citations received during the time period is at least 10 times the mean citation rate in the particular subfield. The particular threshold of selection is somewhat arbitrary. Another definition or set of criteria would give a different sample. Still, the identified HCPs represent the very top papers in their fields.

For MCPs and LCPs, a relative standard is also used. A publication is been considered as low-cited if the number of citations received is less than the mean citation rate of the particular subfield. And the rest publications in this particular subfield are considered as medium-cited.

Identifying the sample data

It is a time-consuming work to gather information from all the papers for establishing papers' feature space. We established a sample set for our experiments. Using the score value of 10 as a selection criterion, about 0.5 % of papers in each journal were identified as HCPs (The *Journal of Mathematical Physics* is an exception, where the ratio of HCPs is about 0.1 %). All the HCPs identified in each journal were contained in the sample set. But, there are large amount of papers were identified as MCPs and LCPs using the selection criteria. Here, only 10 % of MCPs and LCPs were randomly selected out as the sample data for each journal. The distributions of papers in the three levels for the four journals are shown in Table 1.

Establishing the feature space

Both the external and the quality information about these papers were considered in constructing the feature space. The external features mainly come from three aspects: the authors, the journals, and the external features of the paper itself. Sixteen external features were extracted out. The quality features come from papers' knowledge diffusion and first citation properties in the scientific community. Nine features, including papers' citing diffusion properties in the period of 5 years after their publication, and the information associated with papers' first citation were picked up. Table 2 shows the feature space for these papers.

Here, a five-year interval was used to extract papers' quality features. The reason is that it is often used in bibliometric analyses and is intermediate with respect to a short and a long-term citation window. Since the variability of citedness is expected to increase with the size of the citation window, a five-year interval is sufficient long term for a distinct polarization pattern to occur (Aksnes 2003).

There are larger differences in each feature among journals. Each feature in the feature space is normalized first in each journal according to Max–Min normalization method. And every one was transformed into the feature with value located in [0,1]. Then, the normalized data in the four journals are combined into the final sample data for our experiment.

Methods

As it is shown in Table 2, there are twenty-five features in the feature space. The information provided by these features may be reducible. And some noisy samples may also exist to destroy the prediction. The soft fuzzy rough set can help to reduce the feature space by selecting the suitable features, and lessening the negative influences of noises.

Feature selection by soft fuzzy rough set

Here, some definitions for soft fuzzy rough set were given first. More detailed discussion could be found in article (Hu et al. 2010), where the soft fuzzy rough set was first introduced.

The soft fuzzy rough set relies on the soft distance, which is distinguished with the statistical minimum distance in fuzzy rough set (Dubois and Prade 1990).

Table 1 Distributions of papers in each journal

ID	Journals	HCPs	MCPs	LCPs	Total
1	IEEE Transactions on Automatic Control	1	8	17	26
2	Journal of Applied Physics	10	48	133	191
3	Journal of Experimental Medicine	1	9	20	30
4	Journal of Mathematical Physics	5	13	33	51

Table 2 Feature space

Feature	Definition	Source
x_1	Number of authors	External
x_2	Whether there is international cooperation	External
x_3	Whether any author is American	External
x_4	The *h* index of the first author before publication of this paper	External
x_5	The number of papers published by the first author before this paper	External
x_6	The total citations to the papers published by the first author before this paper	External
x_7	The average citations to the paper published by the first author before this paper	External
x_8	The maximum number of papers published by the authors before this paper	External
x_9	The maximum total citations to the papers published by the authors before this paper	External
x_{10}	The maximum *h* index of the authors before publication of this paper	External
x_{11}	The impact factor of the journal publishing this paper	External
x_{12}	The number of papers published in the journal in year 1985	External
x_{13}	The length of this paper	External
x_{14}	The document type of this paper	External
x_{15}	The language of this paper	External
x_{16}	The number of references listed in this paper	External
x_{17}	The first-cited age of this paper	Quality
x_{18}	The first-citations obtained in the first cited year	Quality
x_{19}	The total citations to this paper in its first 5 years after publication	Quality
x_{20}	The number of countries citing this paper in its first 5 years after publication	Quality
x_{21}	The number of document types of papers citing this paper in its first 5 years after publication	Quality
x_{22}	The number of institutions citing this paper in its first 5 years after publication	Quality
x_{23}	The number of languages citing this paper in its first 5 years after publication	Quality
x_{24}	The number of journals citing this paper in its first five years after publication	Quality
x_{25}	The number of subjects citing this paper in its first 5 years after publication	Quality

Definition 1 Given an object x and a set of objects Y, the soft distance between x and Y is defined as:

$$SD\ (x, Y) = \arg_{d(x,y)} \sup_{y \in Y} \{\mathrm{d}(x, y) - \beta m_Y\} \tag{1}$$

The soft distance could help to identify and neglect noisy samples. The penalty factor β is used to control the number of overlooked samples. Parameter $m_Y = |\{y_i|\mathrm{d}(x,\ y_i) < \mathrm{d}(x,\ y)\}|$ shows the number of overlooked samples. If $\mathrm{d}(x,\ y') - \beta m_Y\ (y' \quad Y)$ is the largest of $\{\mathrm{d}(x,\ y) - \beta m_Y\ (\forall y \quad Y)\}$, the distance $\mathrm{d}(x,\ y')$ would be taken as the soft distance between x and Y.

Different from the statistical minimum distance $\mathrm{d}(x,\ Y)$ to calculate the nearest distance of x–Y, the soft distance $SD(x,\ Y)$ is calculated by kth samples, k is controlled by the parameter β. We also discussed the value domain of β, and found that $\beta = 0.1$ is a good choice (Hu et al. 2010). It means that if the soft distance increases 0.1, there's one sample at most is taken as outlier and neglected. In this paper, β is also assigned as 0.1.

Based on the soft distance, the soft fuzzy lower and upper approximations are defined as:

Definition 2 Let U be a nonempty universe, R be a fuzzy similarity relation on U and $F(U)$ be the fuzzy power set of U. The soft fuzzy lower and upper approximations of $A \in F(U)$ are defined as:

$$\begin{cases} \underline{R^S}(A)(x) = 1 - R(x, \arg_y \sup\limits_{A(y) \le A(y_L)} \{1 - R(x,y) - \beta m_{Y_L}\}) \\ \overline{R^S}(A)(x) = R(x, \arg_y \inf\limits_{A(y) \ge A(y_U)} \{R(x,y) + \beta n_{Y_U}\}) \end{cases} \tag{2}$$

where

$$\begin{cases} Y_L = \{y | A(y) \le A(y_L), y \in U\}, y_L = \arg_y \inf\limits_{y \in U} \max\{1 - R(x,y), A(y)\} \\ Y_U = \{y | A(y) \ge A(y_U), y \in U\}, y_U = \arg_y \sup\limits_{y \in U} \min\{R(x,y), A(y)\} \end{cases} \tag{3}$$

m_{Y_L} is the number of the samples overlooked in computing the soft fuzzy lower approximation, $n_{Y_U}$is the number of the samples overlooked in computing the soft fuzzy upper approximation.

Definition 3 Given a decision table $DS = <U, C \cup D>$, U is a nonempty universe, C is the set of attributes and D is the decision attribute. For $\forall B \subseteq C$, the membership of an object $x \in U$ belonging to the soft positive region of D on B is defined as:

$$POS_B^S(D)(x) = \sup_{x \in U/D} \underline{B^S}(X)(x) \tag{4}$$

The soft fuzzy dependency of decision D on feature subset B is defined as:

$$\gamma_B^S(D) = \frac{\sum_{x \in U} POS_B^S(D)(x)}{|U|} \tag{5}$$

Dependency is the ratio of the samples in the lower approximation over the universe, which is widely used to measure the classification performance of attributes. A larger dependency of feature subset means that it has better capability to distinguish different classes.

Based on the soft fuzzy dependency, we designed a feature selection algorithm to select features that have influences on papers' citation impact. Table 3 shows the feature selection algorithm. The algorithm employs the soft fuzzy dependency as the feature evaluation function and the sequential forward selection as the search strategy. The output of the algorithm is a feature ranking set $F' = \{f_1', f_2', \ldots, f_{|F''|}'\}$. Given the set F_{k-1}' with $k - 1$ features selected, the k'th feature is determined by $\max\limits_{f \in F - F_{K-1}'} \{SFD_{\{F_{K-1}' \cup \{f\}\}}(D)\}$. Thus, the feature with the maximum soft fuzzy dependency is extracted in every circulation. Finally, a sequential of feature subsets of $F_1' = \{f_1'\}$, $F_2' = \{f_1', f_2'\}, \ldots,$ $F_{F'}' = \{f_1', f_2', \ldots, f_{|F''|}'\}$ is generated.

Then, the KNN classifier is used to cross-validate the classification accuracy of the feature subsets.

Classification by KNN classifier

The KNN algorithm is amongst the simplest of all machine learning algorithms for classifying objects based on the closest k training examples in the feature space (Cover and

Table 3 Feature selection algorithm

Input	X, F	X is a sample set and F is the original feature set
Output	F'	F' is a ranking feature set
	Begin	
	Initialize	$F' = \varphi$
	While	$F \neq \varphi$
		Find $f = \arg_f \max_{f \in F} \{SFD_{\{F' \cup \{f\}\}}(D)\}$
		$F' = F' \cup \{f\}$
		$F = F - \{f\}$
	End	
	Return	F'
	End	

Hart 1967). An object would be assigned to the class most common amongst its k nearest neighbors.

In this paper, the KNN classifier ($k = 3$) is used to cross-validate the classification accuracy of the data with these feature subsets. The feature subset with the highest classification accuracy would be chosen out as the final predictors.

Results

Performing the feature selection algorithm as shown in Table 3, we obtain thirteen feature subsets from the feature space. The result is shown in Table 4 as rank of 1–13. Here, there are two questions which must be addressed in any attempt to use these subsets as predictors.

> Question I: Whether the subsets selected are capable of discriminating between different papers and, by extension, whether there's some subset is the best.
>
> Question II: Whether the subsets chosen can reflect the factors that the researchers like to probe.

For Question I, the KNN classifier ($k = 3$) is used to cross-validate the classification accuracy of these feature subsets. The feature subset with the highest classification accuracy would be chosen out as the final predictors. The last column in Table 4 shows the classification performance of these subsets. The last row in Table 4 also shows the classification result on the original feature space. Obviously, almost all of the thirteen feature subsets obtain better classification performances than the original one, except for the subset of $\{x_4\}$. And the feature subset of $\{x_4, x_{24}, x_{25}, x_{17}, x_{19}, x_{20}, x_{22}\}$, the seventh row written in bold, could most accurately predict whether a paper would grow into a HCP, MCP or LCP in future. It indicates that among the twenty-five features in the original feature space, seven features could be competent to determine papers' future citation impact.

For Question II, a feature-specific analysis was performed to explore whether the selected features are the factors that researchers like to probe.

(i) The h index of the first author $\{x_4\}$ before publication of this paper: The h-index is an index proposed by Hirsch (2005) to evaluate the quality of scientific research from a micro viewpoint. A larger h-index indicates that an author has gained considerable

Table 4 Selected feature subsets and classification accuracies

Rank	The feature subsets $\{x_i\}$	Classification accuracies by KNN
1	$\{x_4\}$	0.7143
2	$\{x_4, x_{24}\}$	0.7619
3	$\{x_4, x_{24}, x_{25}\}$	0.788
4	$\{x_4, x_{24}, x_{25}, x_{17}\}$	0.8083
5	$\{x_4, x_{24}, x_{25}, x_{17}, x_{19}\}$	0.8205
6	$\{x_4, x_{24}, x_{25}, x_{17}, x_{19}, x_{20}\}$	0.8571
7	$\boldsymbol{\{x_4, x_{24}, x_{25}, x_{17}, x_{19}, x_{20}, x_{22}\}}$	**0.9048**
8	$\{x_4, x_{24}, x_{25}, x_{17}, x_{19}, x_{20}, \boldsymbol{x_{22}}, x_{10}\}$	0.880
9	$\{x_4, x_{24}, x_{25}, x_{17}, x_{19}, x_{20}, \boldsymbol{x_{22}}, x_{10}, x_{18}\}$	0.8571
10	$\{x_4, x_{24}, x_{25}, x_{17}, x_{19}, x_{20}, \boldsymbol{x_{22}}, x_{10}, x_{18}, x_{14}\}$	0.838
11	$\{x_4, x_{24}, x_{25}, x_{17}, x_{19}, x_{20}, \boldsymbol{x_{22}}, x_{10}, x_{18}, x_{14}, x_2\}$	0.8143
12	$\{x_4, x_{24}, x_{25}, x_{17}, x_{19}, x_{20}, \boldsymbol{x_{22}}, x_{10}, x_{18}, x_{14}, x_2, x_5\}$	0.8095
13	$\{x_4, x_{24}, x_{25}, x_{17}, x_{19}, x_{20}, \boldsymbol{x_{22}}, x_{10}, x_{18}, x_{14}, x_2, x_5, x_8\}$	0.8023
***	The original feature space of $\{x_1, x_2,\ldots, x_{25}\}$	0.7396

research capabilities or reputations in science. Van Dalen and Henkens (2001, 2005) suggested that authors with high reputations could receive disproportionately more citations than authors with low reputations. Levitt and Thelwall (2009) investigated the citation profiles of first author based on 82 most highly cited *Information Science and Library Science* (IS&LS) papers. They found that high percentage of the first authors with relative lower *h* index in the field of IS&LS, but they also showed that some of them may be highly cited in other fields.

(ii) The citation properties in the first 5 years after a paper's publication $\{x_{24}, x_{25}, x_{19}, x_{20}, x_{22}\}$: These five features represent papers' knowledge diffusion properties in the scientific community, which were also extracted out to be the typical features associated with papers' citation counts in our preliminary work (Wang et al. 2011). The wider citation distributions in various journals, subjects, countries and institutions increase a paper's visibility to a larger extent, and so to their citation counts.

(iii) The first-cited age $\{x_{17}\}$ of one paper: The first-cited age shows the rate of a paper to be accepted after publication. It indicates that the easier a paper is cited, the larger the probability of it being cited frequently. Van Dalen and Henkens (2005) showed that the status of uncitedness of a paper becomes a stigma and the longer a paper is uncited, the lower its quality and the less inclined researchers will be to cite it. Glänzel et al. (2003) stated that the probability of a paper's uncitedness increases dramatically with belated first citations, and the probability of being frequently cited later on decreases to the same extent. In fact, the negative influences brought about by a paper's uncitedness reflect the importance of the position of a paper's first citation on its later citation life.

Therefore, our results show that it is feasible to early recognition papers' future citation impact by objectively assessed factors. Compared with the prior endeavors to seek the possible factors influencing on papers' citation impact, we not only consider the external information about the journal and paper and authors, but also the information about papers' quality. And we showed that both the papers' quality and papers' external features, mainly represented as the reputation of the first author, are contributed to papers' future citation impact.

These features show some intriguing patterns in HCPs. The concept of quality and visibility dynamics can be useful in order to understand some of the mechanisms involved.

High citation scores are the results of many researchers' decisions to cited a particular paper. Though there may be a large number of reasons why an author cites a particular paper, the concept of cognitive differentiation is still relevant because scientists tend to cite papers that are useful for their own research. The sign of paper: (a) being noted immediately; and (b) being cited widely by the scientific community, is an informative signal for their judgment. This is the quality dynamics.

By the visibility dynamics, we mean that certain social mechanisms can influence citation rates. As we have shown that the reputation of author plays a role in gaining attention. This is a variant of the "Matthew effect" (Merton 1968), stating that recognition is skewed in favor of established scientists. When a paper is published by a well-known person, even more people will become aware of it. Its visibility increases and thereby the chances of getting even more citations. Similarly, when a paper has received many citations, the paper obtains status as an authoritative paper. Consequently, even more researchers will cite it, because appealing to existing authorities may be one reason for citing a paper.

Conclusions

In summary, our results suggest that the papers' citation impact could be predicted by objective bibliometric factors. Both the variables about papers' external bibliometric features and the variables about papers' quality were considered in constructing papers' feature space. Because the information provided by these features may be reducible and some noisy samples may also exist to destroy the prediction, the soft fuzzy rough set was used to reduce the feature space by selecting the suitable features, and lessening the negative influences of noises. Then, the KNN classifier ($k = 3$) was used to examine the reduced feature subset. The feature subset with the highest classification accuracy is extracted, having the key features influencing on papers' future citation impact. And those following factors were proved to be the key features:

(a) The research capabilities of the first author, represented by his/her h index before the publication of this paper;
(b) The papers' quality, represented by the papers' earlier knowledge diffusion properties and the first-cited properties in the scientific community.

Something should be mentioned for the feature of $\{x_{11}\}$, the impact factor of one journal. Because only one journal in one field is considered to be the data source, this feature wasn't identified as the core indicator. But it did not conceal the significant role of the journal on papers' citation impact. In our preliminary work (Wang et al. 2011), we found that the reputation of journals plays an overriding role in gaining attention in science. The similar conclusions have also been suggested by Van Dalen and Henkens (2001, 2005).

Acknowledgments This work was supported by the National Natural Science Foundation of China (Grant Nos. 71003020; 70973031), the special funds of Central College Basic Scientific Research Bursary (Grant No. DL11CB09), and the Postdoctoral Science Foundation of Heilongjiang Province.

References

Aksnes, D. W. (2003). Characteristics of highly cited papers. *Research Evaluation, 12*(3), 159–170.

Bornmann, L., & Daniel, H. D. (2008). What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation, 64*(1), 45–80.

Burrell, Q. L. (2001). Stochastic modelling of the first-citation distribution. *Scientometrics, 52*, 3–12.

Burrell, Q. L. (2002a). On the nth-citation distribution and obsolescence. *Scientometrics, 53*, 309–323.

Burrell, Q. L. (2002b). Will this paper ever be cited? *Journal of the American Society for Information Science and Technology, 53*, 232–235.

Burrell, Q. L. (2003). Predicting future citation behavior. *Journal of the American Society for Information Science and Technology, 54*(5), 372–378.

Case, D. O., & Higgins, G. M. (2000). How can we investigate citation behavior? A study of reasons for citing literature in communication. *Journal of the American Society for Information Science, 51*(7), 635–645.

Cover, T. M., & Hart, P. E. (1967). Nearest neighbor pattern classification. *IEEE Transaction on Information Theory, IT-13*(1), 21–27.

Danell, R. (2011). Can the quality of scientific work be predicted using information on the author's track record? *Journal of the American Society for Information Science and Technology, 62*(1), 50–60.

Dubois, D., & Prade, H. (1990). Rough fuzzy sets and fuzzy rough sets. *General Systems, 17*, 191–209.

Fu, L., & Aliferis, C. (2010). Using content-based and bibliometric features for machine learning models to predict citation counts in the biomedical literature. *Scientometrics, 85*, 257–270.

Glänzel, W. (1997). On the reliability of predictions based on stochastic citation processes. *Scientometrics, 40*(3), 481–492.

Glänzel, W., Rinia, E. J., & Brocken, M. G. M. (1995). A bibliometric study of highly cited European physics papers in the 80s. *Research Evaluation, 5*(2), 113–122.

Glänzel, W., & Schubert, A. (1995). Predictive aspects of a stochastic model for citation processes. *Information Processing and Management, 31*(1), 69–80.

Glänzel, W., Schlemmer, B., & Thijs, B. (2003). Better late than never? On the chance to become highly cited only beyond the standard bibliometric time horizon. *Scientometrics, 58*(3), 571–586.

Hewings, A., Lillis, T., & Vladimirou, D. (2010). Who's citing whose writings? A corpus based study of citations as interpersonal resource in English medium national and English medium international journals. *Journal of English for Academic Purposes, 9*(2), 102–115.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the USA, 102*(46), 16569–16572.

Hu, Q. H., An, S., & Yu, D. R. (2010). Soft fuzzy rough sets for robust feature evaluation and selection. *Information Sciences, 180*, 4384–4400.

Kim, K. (2004). The motivation for citing specific references by social scientists in Korea: The phenomenon of co-existing references. *Scientometrics, 59*(1), 79–93.

Laband, D. N., & Piette, M. J. (1994). Favoritism versus search for good papers: Empirical evidence regarding the behavior of journal editors. *Journal of Political Economy, 102*, 194–203.

Levitt, J. M., & Thelwall, M. (2008). Is multidisciplinary research more highly cited? A macrolevel study. *Journal of the American Society for Information Science and Technology, 59*(12), 1973–1984.

Levitt, J. M., & Thelwall, M. (2009). The most highly cited Library and Information Science articles: Interdisciplinarity, first authors and citation patterns. *Scientometrics, 78*(1), 45–67.

Merton, R. K. (1968). The Matthew effect in science. *Science, 159*, 56–63.

Penas, C. S., & Willett, P. (2006). Gender differences in publication and citation counts in librarianship and information science research. *Journal of Information Science, 32*(5), 480–485.

Rong, T., & Martin, A. S. (2008). Author-rated importance of cited references in biology and psychology publications. *Journal of Documentation, 64*(2), 246–272.

Sagi, I., & Yechiam, E. (2008). Amusing titles in scientific journals and article citation. *Journal of Information Science, 34*(5), 680–687.

Van Dalen, H. P., & Henkens, K. (2001). What makes a scientific article influential? The case of demographers. *Scientometrics, 50*, 455–482.

Van Dalen, H. P., & Henkens, K. (2005). Signals in science-on the importance of signaling in gaining attention in science. *Scientometrics, 64*(2), 209–233.

Wang, M. Y., Yu, G., & Yu, D. R. (2011). Mining typical features for highly cited papers. *Scientometrics, 87*(3), 695–706.

Xia, J. F., Myers, R. L., & Wihoite, S. K. (2011). Multiple open access availability and citation impact. *Journal of Information Science, 37*(1), 19–28.